MRR-FV: Unlocking Complex Fact Verification with Multi-Hop Retrieval and Reasoning

Liwen Zheng[1], Chaozhuo Li[1]*, Litian Zhang[2], Haoran Jia[1], Senzhang Wang[3], Zheng Liu[4], Xi Zhang[1]

[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE), Beijing University of Posts and Telecommunications, China
[2]Beijing University of Aeronautics and Astronautics, Beijing 100191, China
[3]Central South University, China
[4]BAAI, China
{zhenglw, lichaozhuo}@bupt.edu.cn, litianzhang@buaa.edu.cn, jiahaoran@bupt.edu.cn, szwang@csu.edu.cn, zhengliu1026@gmail.com, zhangx@bupt.edu.cn

Abstract

The pervasive spread of misinformation on social networks highlights the critical necessity for effective fact verification systems. Traditional approaches primarily focus on pairwise correlations between claims and evidence, often neglecting comprehensive multi-hop retrieval and reasoning, which results in suboptimal performance when dealing with complex claims. In this paper, we propose MRR-FV, a generative retrieval-enhanced model designed to address the novel challenge of Multi-hop Retrieval and Reasoning for Fact Verification, which integrates two core modules: Generative Multi-hop Retriever and the Hierarchical Interaction Reasoner. MRR-FV utilizes an autoregressive model for iterative multi-hop evidence retrieval, complemented by a pre-trained compressor to address the challenge of intention shift across retrieval hops. For claim verification, we propose a hierarchical interaction reasoner that conducts intra-sentence reasoning to capture long-term semantic dependencies and inter-sentence reasoning across multi-hop evidence subgraphs to reveal complex evidence interactions. Experimental evaluations on the FEVER and HOVER datasets demonstrate the superior performance of our model in both claim verification and evidence retrieval tasks.

Introduction

Fact verification (FV) aims to leverage credible evidence to autonomously assess the veracity of textual statements, which helps combat the proliferation of misinformation and enhance the reliability and credibility of social media (Guo, Schlichtkrull, and Vlachos 2022; Zhang et al. 2025). Existing FV models usually follow a two-phase paradigm consisting of evidence retrieval and claim verification (Hu et al. 2023). Evidence retrieval focuses on precisely identifying critical evidential sentences within a vast corpus (Chen et al. 2022a). In the claim verification phase, the semantic interactions between the claim and the retrieved evidence are organized into sequences or graphs, which are then jointly analyzed to assess their authenticity (Zeng, Abumansour, and Zubiaga 2021).

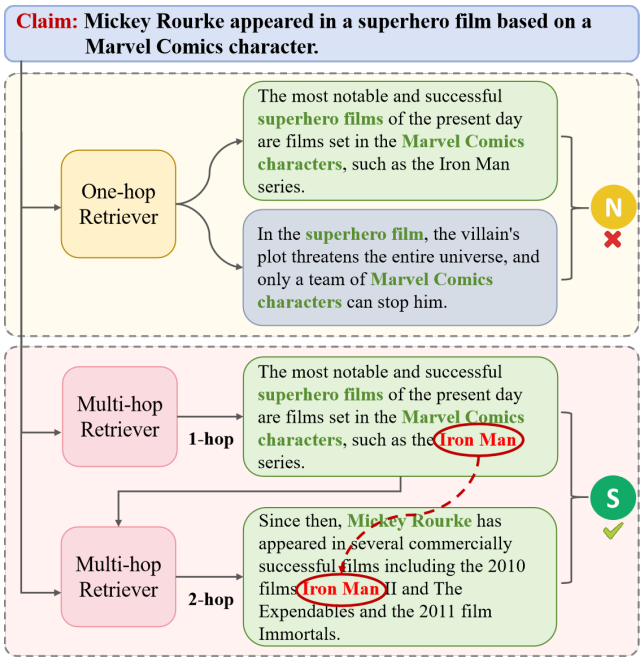


Figure 1: The evidence retrieved by the one-hop retriever leads to an incorrect verification label of "Not Enough Information," indicated as "N" in the figure above. However, the evidence retrieved by the multi-hop retriever can result in the correct verification label of "Supports," denoted as "S".

Evidence retrieval serves as the cornerstone of FV systems, offering a robust and essential foundation for the subsequent claim verification (Samarinas, Hsu, and Lee 2021). Most retrieval modules adhere to a one-hop retrieval paradigm, where the semantic similarity between each candidate evidence and the claim is assessed independently. This approach focuses solely on the semantic relationship within individual claim-evidence pairs, overlooking the potential interactions between the claim and a wider set of evidence (Li et al. 2021). Consequently, this limitation may result in sub-optimal performance when addressing complex

*Corresponding author: Chaozhuo Li.

claims that necessitate support from multiple pieces of evidence. As depicted in Figure 1, the one-hop retriever identifies two evidence that are semantically aligned with the claim. However, both pieces of evidence fail to capture the critical information concerning "Mickey Rourke", rendering the combination of these independently retrieved evidence insufficient for verifying the claim.

Moving beyond the traditional one-hop retrieval strategy, we emphasize fact verification enhanced by multi-hop evidence retrieval, where evidence is selected not only based on the claim but also by leveraging information accumulated from previous hops (Liao et al. 2023; Chen et al. 2024). As illustrated in Figure1, the two-hop evidence is selected by considering both the input claim and the first-hop evidence. While the two-hop evidence may be less directly aligned with the claim, it incorporates crucial information (e.g., "Iron Man") from the one-hop evidence, thereby providing complementary details that contribute to a more comprehensive set of evidence to support the claim.

Few efforts have explored integrating multi-hop evidence retrieval into FV systems (Subramanian and Lee 2020). The common approach involves iterative dense retrieval, where evidence from previous hops is concatenated with the current query for subsequent retrievals (Zhu et al. 2023; Zhang et al. 2023). However, as more evidence is added, irrelevant information may accumulate, causing the final query to drift semantically from the claim, which may undermine verification accuracy (Zhang et al. 2024b). To address the limitations, our motivation lies in designing generative paradigm for multi-hop evidence retrieval. Unlike traditional discriminative methods, generative approaches can dynamically generate next-hop evidence in an autoregressive manner, introducing constraints to reduce intention shift instead of merely concatenating prior evidence with the original claim.

The generative multi-hop retrieval mechanism effectively aggregates comprehensive evidence across multiple hops. Nevertheless, the intricate interrelationships among the retrieved evidence introduce substantial complexity to the ensuing multi-hop reasoning process within the claim verification phase (Zhang et al. 2024c,a). Existing approaches to multi-hop reasoning predominantly employ either sequence-based or graph-based architectures (Guo, Schlichtkrull, and Vlachos 2022). Sequence-based methods are particularly effective in capturing local contextual information (Zhao et al. 2021), while graph-based methods excel in handling multiple evidence nodes and their complex interconnections, facilitating the synthesis of global information (Lan et al. 2024; Li et al. 2017). Considering the complementary roles of local and global information in multi-hop reasoning (Zhang, Zhang, and Pan 2022), it is imperative to develop a unified reasoning module capable of capturing both hierarchical semantics and multi-hop structural features.

In this paper, we propose a novel **MRR-FV** model to alleviate the problem of **M**ulti-hop **R**etrieval and **R**easoning for **F**act **V**erification. In the evidence retrieval phase, the autoregressive generative model BART (Lewis et al. 2020) is employed as the backbone, supervised and optimized through joint training with the claim verification module. To address intention shift in multi-hop retrieval, we propose a novel query compressor that condenses the claim and historical evidence into a query. To enhance the compressor's generation capability, we introduce a novel pre-training strategy that integrates the powerful generative capacity of large language models (LLMs) into smaller language models (SLMs). The compressed query is then input into the constrained generative retriever to generate evidence from the candidate pool. For claim verification, we propose a hierarchical interaction reasoner to simultaneously encode intra-sentence reasoning and inter-sentence correlations, integrating long-term semantic dependencies and interactions among multi-hop evidence within a unified reasoning framework. Experimental results on two datasets demonstrate the superiority of our approach. The main contributions of this work include:

- We investigate the novel problems of intention shift in multi-hop retrieval and explicit multi-hop reasoning in verification by modeling deep interactive features between evidence throughout the retrieval and verification.
- We further propose MRR-FV, a generative retrieval-enhanced model with two key modules: the Generative Multi-hop Retriever for constructing a complete and coherent evidence set, and the Hierarchical Interaction Reasoner, which sequentially extracts critical clues from intra and inter-sentence to facilitate claim verification.
- We conduct extensive experiments on both the FEVER and HOVER datasets, demonstrating superior performance and underscoring the effectiveness of MRR-FV.

Related Work

Fact verification involves assessing the veracity of a given claim by validating it against reliable evidence (Guo, Schlichtkrull, and Vlachos 2022). At the evidence retrieval stage, one-hop methods account for the relationship between claim and evidence (Fajcik, Motlicek, and Smrz 2023a). However, their lack of interactive capabilities during retrieval limits the effectiveness (Liao et al. 2023). Multi-hop retrieval mechanisms address this limitation by enabling iterative retrieval and integration of multiple pieces of evidence, thereby enhancing the system's capacity to manage complex claims (Xiao et al. 2024). However, current multi-hop methods remain constrained by their heavy reliance on hyperlinks (Subramanian and Lee 2020) and are prone to intention shift, limiting their overall effectiveness (Khattab, Potts, and Zaharia 2021).

During the stage of claim verification, existing approaches predominantly emphasize claim verification through the exploration of fine-grained reasoning methods based on sequential (Wu, Wang, and Zhao 2024) or graph (Xu et al. 2022; Zhang, Zhang, and Zhou 2024) structures to integrate existing evidence. However, the absence of a unified reasoning framework capable of concurrently modeling both intra-sentence contextual features and inter-sentence structural relationships presents significant challenges in addressing the complex interactions inherent in multi-hop evidence. Recent advancements utilize the reasoning capabilities of large language models, achieving notable performance improvements (Zhang and Gao 2023; Yue et al. 2024; Liu et al.

2024). Nevertheless, these approaches often demand substantial resource integration and tend to neglect fine-grained evidence retrieval, limiting their overall effectiveness.

Problem Definition

Fact verification evaluates the truthfulness of a claim c based on an evidence set $E = \{e_1, e_2, \ldots, e_n\}$ sourced from a large corpus of documents D to either support or refute the claim, in which e_i denotes a single evidence. The annotation y can be categorized as "Supports" or "Refutes" and occasionally as "Not Enough Information". This process requires not only a precise understanding of the relationship between the claim and the evidence but also efficient retrieval of relevant information from a vast evidence collection.

Methodology

As illustrated in Figure 2, the proposed MRR-FV model consists of two modules: the Generative Multi-hop Retriever, which adeptly retrieves comprehensive evidence by modeling semantic correlations across evidence, and the Hierarchical Interaction Reasoner, which captures long-term dependencies within individual sentences as well as multi-hop interactions across multiple sentences.

Generative Multi-hop Retriever

Given the claim c, the generative multi-hop retriever aims to iteratively select a set of evidence $E_{1:n} = \{e_1, e_2, \ldots, e_n\}$ using generative models. Specifically, at each hop t, the query compressor encodes the sequence formed by concatenating the claim c with the previously retrieved evidence $E_{1:t-1}$, generating a compressed query Q_t that maintains semantic consistency with the original claim. The constrained generative retriever then utilizes a seq2seq model to retrieve the next-hop evidence e_t from the corpus based on Q_t. This iterative process entails leveraging historical evidence to refine subsequent retrievals at each step, ultimately synthesizing a comprehensive chain of evidence to facilitate robust claim verification.

Query Compressor During retrieval, directly concatenating the claim with multi-hop evidence can make input queries excessively long, diminishing the claim's semantic integrity and impairing evidence retrieval accuracy. To address this, we introduce a query compressor that condenses the extended text while retaining the core semantics and relationships between the claim and evidence.

At the t-th hop, as illustrated in Figure 2(a), the query compressor utilizes the original claim c and evidence retrieved in the previous $t-1$ hops, denoted as $E_{1:t-1} = \{e_1, e_2, \ldots, e_{t-1}\}$, as input, and generates the compressed query Q_t for the t-th evidence retrieval. Specifically, given the input query $I_t = \{c; e_1; e_2; \ldots; e_{t-1}\}$, where $[;]$ denotes the concatenation operation, the encoder first processes I_t into a contextual representation $h_t = \text{Encoder}_C(I_t)$. At each decoding step j, the decoder generates the j-th token $q_{t,j}$ of Q_t based on h_t and the previously generated tokens $q_{t,<j}$:

$$\mathbb{P}(q_{t,j} \mid q_{t,<j}, h_t) = \text{Decoder}_C(q_{t,<j}, h_t), \quad j \leq L, \quad (1)$$

where L is the pre-defined maximum compression length.

Enhanced Pre-training of the Query Compressor To mitigate intention shift and construct a more compact framework, we leverage the powerful inductive capabilities of LLMs to pre-train the query compressor, as depicted in Figure 2(b). (1) **Distillation** to endow the query compressor with compression ability. Given the claim c and the evidence list E, we ask the LLM to generate the compressed text Q', which serves as the learning target for the query compressor given c and E. (2) **Alignment** to endow the query compressor with intention consistency. Given c and E, we use an LLM to assess the matching degree between N compressed queries and c, designating the highest-scoring query as the positive sample Q^+ and the lowest-scoring queries as negatives $\{Q_1^-, Q_2^-, \ldots, Q_n^-\}$. Subsequently, we apply an InfoNCE loss L_a to align the compressed queries with Q^+:

$$L_a = -\frac{1}{N}\sum_{i=1}^{N} \log \left(\frac{\exp\left(\frac{f(Q_i)\cdot f(Q^+)}{\tau}\right)}{\sum_{j=1}^{n} \exp\left(\frac{f(Q_i)\cdot f(Q_j)}{\tau}\right)} \right) \quad (2)$$

where $f(\cdot)$ is the encoder and τ is the temperature coefficient. The pre-training process ensures that compressed queries align strongly with the original claim, optimizing compression to mitigate intention shift and enhance fact verification.

Constrained Generative Retriever During the retrieval process, the interaction between evidence is enhanced by employing an autoregressive generation approach. At each step, the generative retriever takes the original query and the historical evidence as input, using a seq2seq model to generate the subsequent evidence.

Initially, the query compressor compresses the input sequence I_t into the compressed query Q_t. During the generation process, the compressed query Q_t is encoded into the contextual representation $h_t^q = \text{Encoder}_G(Q_t)$. Subsequently, the t-th evidence is progressively generated. At each decoding time step j, Decoder_G produces the current token $e_{t,j}$ based on the previously generated tokens $e_{t,<j}$ and h_t^q:

$$\mathbb{P}(e_{t,j} \mid e_{t,<j}, h_t^q) = \text{Decoder}_G(e_{t,<j}, h_t^q). \quad (3)$$

The overall generation process of the t-th evidence e_t can be expressed as:

$$\mathbb{P}(e_t|h_t^q) = \prod_{j=1}^{L} P(e_{t,j}|e_{t,<j}, h_t^q), \quad (4)$$

where L represents the length of e_t, and $\mathbb{P}(e_t|h_t^q)$ denotes the probability distribution of generating e_t given the input contextual representation h_t^q.

Moreover, our goal is to retrieve crucial evidence from the candidate corpus to support claim verification. However, allowing the generation model to select any token from the entire vocabulary at each decoding step may generate evidence that doesn't match valid candidates (Cao et al. 2021). To prevent this issue, we employ a constrained generation strategy restricting the model to decode only valid tokens.

The prefix tree, or trie, is a tree-like data structure used to store strings, with each node representing a character's token id. The path from the node to any node represents a string or

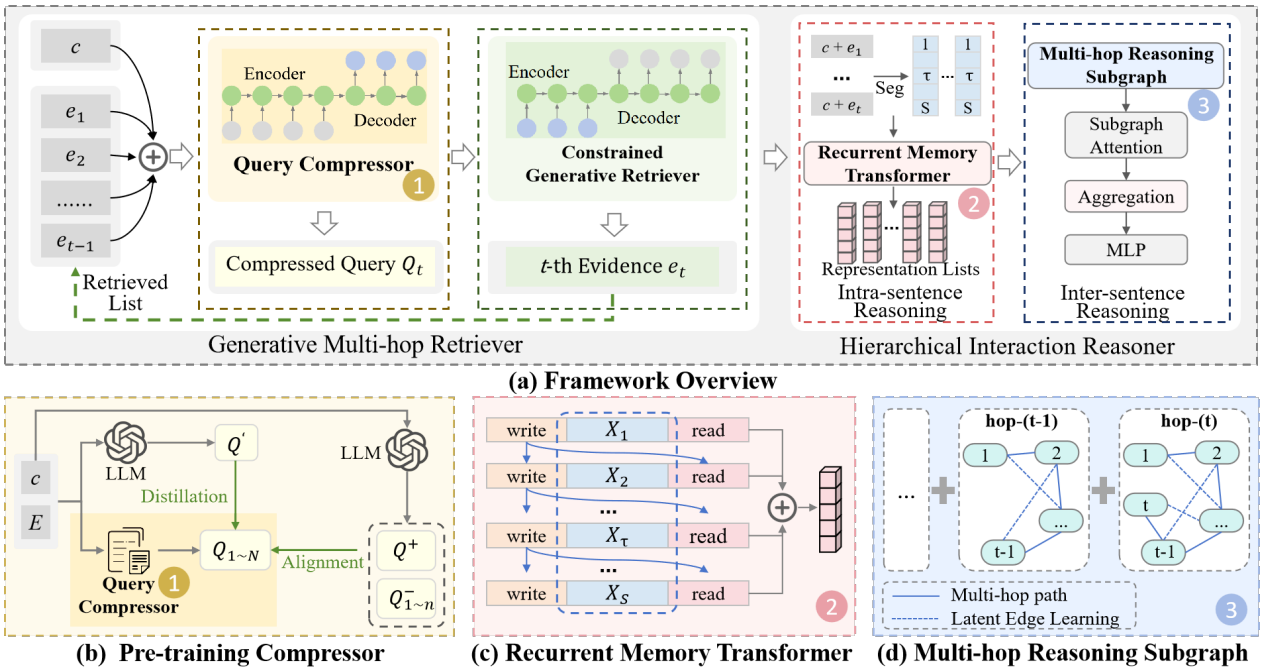


Figure 2: Framework of the proposed MRR-FV model.

its prefix. We store the candidate evidence in a prefix tree to restrict allowed tokens at each decoding step. The prefix tree-based constraint significantly reduces the number of candidate tokens, thereby enhancing generation efficiency.

Hierarchical Interaction Reasoner

To model the clue chain in the multi-hop evidence for claim verification, we conduct reasoning on both sequence and graph structures within a unified framework. This process is divided into two main components: Intra-sentence Reasoning and Inter-sentence Reasoning.

Intra-sentence Reasoning We conduct intra-sentence reasoning on the sequence of each claim-evidence pair to initialize the contextual features. As the number of retrieving hops increases, the distance between the retrieved evidence and the original claim expands. Traditional transformers often blur global information when processing long sequences due to dispersed context. To address this issue, we propose utilizing the Recurrent Memory Transformer (Bulatov, Kuratov, and Burtsev 2022), which incorporates memory tokens to facilitate the transmission and retention of global information across multiple segments, thereby enhancing the capture of global context in long sequences.

As depicted in Figure 2(c), suppose X is the sequence of a claim-evidence pair, which is split into S segments, i.e., $X = [X_1, X_2, \ldots, X_S]$. For effective long-term dependency modeling, the τ-th segment X_τ is augmented with special memory tokens H_τ^{mem}, and the augmented sequence is then processed with a standard Transformer. Specifically, the memory tokens are appended to both the beginning and end of the segment tokens' representations H_τ^0:

$$\begin{aligned} \tilde{H}_\tau^0 &= [H_\tau^{\text{mem}} \circ H_\tau^0 \circ H_\tau^{\text{mem}}], \\ \bar{H}_\tau^N &= \text{Transformer}(\tilde{H}_\tau^0), \\ [H^{\text{read}} \circ H_\tau^N \circ H_\tau^{\text{write}}] &:= \bar{H}_\tau^N, \end{aligned} \tag{5}$$

where N represents the number of Transformer layers. Additionally, to enable the sequence representation to attend to the memory states produced in this segment, the starting group of tokens in $\tilde{H}_\tau^N$ functions as the read memory H_τ^{read}. To attend to all current segment tokens and update representation stored in the memory, the ending group of tokens in $\tilde{H}_\tau^N$ works as the write memory H_τ^{write}, containing updated memory tokens for the τ-th segment.

The $(\tau+1)$-th segment of the input sequence is processed in order. To establish a recurrent connection between segments, the memory tokens output from the current segment are transferred to the input of the subsequent segment:

$$\begin{aligned} H_{\tau+1}^{\text{mem}} &:= H_\tau^{\text{write}}, \\ \tilde{H}_{\tau+1}^0 &= [H_{\tau+1}^{\text{mem}} \circ H_{\tau+1}^0 \circ H_{\tau+1}^{\text{mem}}]. \end{aligned} \tag{6}$$

Finally, the sentence embedding H can be obtained by combining the read memory outputs of each segment. This recurrent memory process efficiently models long-term dependencies and ensures precise information transmission.

Inter-sentence Reasoning As illustrated in Figure 2(d), to capture the semantic connections between the claim and evidence, we construct the evidence graph based on the retrieval path, denoting the t_{th} hop evidence graph as G^t. To mitigate the limitations of sparse graph structures on node feature learning, we leverage latent edge learning to infer potential

connections, enhancing the representation of claim-evidence pairs (Zheng et al. 2022; Li et al. 2018). Following the network's homogeneity principle, where similar nodes are more likely to establish mutual connections, we infer latent edges by computing feature similarity between nodes.

Supposing $H^t = [h_1^t, \ldots, h_{t-1}^t, h_t^t]$ is the initial representation of G^t, where h_i^t denotes the i_{th} nodes in G^t. To integrate semantic and structural features in latent edge learning, we first employ graph convolutional networks to derive the structural feature $h_i^{t-\text{G}}$. This is then combined with the initial semantic representation h_i^t to calculate the similarity between nodes, which is subsequently used to construct the latent edge matrix:

$$h_i^{t-\text{sim}} = \frac{h_i^t + h_i^{t-\text{G}}}{2}. \tag{7}$$

Subsequently, we compute the cosine similarity β_{ij}^t between each pair of nodes and construct latent edges e_{ij}^t using a predetermined similarity threshold γ:

$$\beta_{ij}^t = \frac{h_i^{t-\text{sim}} \cdot h_j^{t-\text{sim}}}{\|h_i^{t-\text{sim}}\| \|h_j^{t-\text{sim}}\|}, \tag{8}$$

$$e_{ij}^t = \begin{cases} 1, & \text{if } \beta_{ij}^t > \gamma \\ 0, & \text{otherwise} \end{cases}. \tag{9}$$

To account for the dynamic nature of node relationships under different connections, we employ a hop-specific trainable weight matrix W_R^t, enabling the calculation of the edge weight matrix E^t for the t-hop evidence graph, where e_{ij}^t represents the relationship value between node i and node j:

$$e_{ij}^t = \text{LeakyReLU}\left(a^\top [W_R^t h_i^t \parallel W_R^t h_j^t]\right), \tag{10}$$

Furthermore, shorter graph distances indicate stronger correlations between nodes. Thus, we introduce graph-relative positions as an attention bias during aggregation. This bias is incorporated into the attention mechanism based on the *softmax* function, thereby enhancing attention between neighboring evidence nodes. The output representation of G^t, denoted as $\tilde{H}^t = S^t H^t$, where s_{ij}^t in S^t is defined as:

$$s_{ij}^t = \frac{\exp(e_{ij}^t - m_G |d_{ij}|)}{\sum_{k=1}^n \exp(e_{ik}^t - m_G |d_{ik}|)}, \tag{11}$$

where d_{ij} denotes the graph-relative distance between evidence nodes i and j, and m_G represents the fixed slope for a specific head. To stabilize the learning process, the aforementioned mechanism can be extended to a multi-head format with h heads. After concatenating the outputs of each head, the final graph representation $\tilde{H}^t$ can be obtained through a normalization layer:

$$\tilde{H}^t = \text{Norm}(\text{concat}(\tilde{H}_{(1)}^t, \tilde{H}_{(2)}^t, \ldots, \tilde{H}_{(h)}^t)). \tag{12}$$

Subsequently, we derive the final fusion representation F by concatenating the multi-hop representations:

$$F^t = \text{concat}(\tilde{H}^1, \tilde{H}^2, \ldots, \tilde{H}^t). \tag{13}$$

Ultimately, we feed the resulting representation F^t into MLP to obtain the predicted label.

Joint Training Mechanism

To oversee the intermediate stages of multi-hop retrieval, we verify the claim following each retrieval step to assess the current evidence chain. We compute the verification loss L_v using the verification outcomes, and proceed with joint training of the retriever and reasoner through gradient back propagation. Furthermore, we calculate the retrieval loss L_r by assessing whether the retrieved evidence is contained within the gold evidence set, offering additional supervision to enhance the effectiveness of the retrieval process.

Experiments

Experimental Setup

This section describes the dataset, evaluation metrics, and baselines of our experiments.

Datasets We conduct our evaluations using the large-scale dataset FEVER (Thorne et al. 2018) and HOVER (Jiang et al. 2020), a multi-hop fact-verification dataset. FEVER comprises 185,455 annotated claims alongside 5,416,537 Wikipedia documents. HOVER includes claims that necessitate integration and reasoning across multiple Wikipedia articles, with claims verifiable in 2-4 hops. Besides, we use the dev set of HOVER for evaluation since the test sets are not publicly released. All claims in both FEVER and HOVER are classified by annotators as Supports, Refutes, or Not Enough Info.

Evaluation Metrics Following previous studies, we employ Label Accuracy (LA) and the FEVER score as evaluation metrics for claim verification on the FEVER dataset (Hanselowski et al. 2018; Liu et al. 2020). LA is a widely applicable metric, and the FEVER score considers label accuracy contingent upon the provision of at least one complete set of golden evidence (Thorne et al. 2018). Additionally, for evidence retrieval, given the gold evidence, we utilize Precision, Recall, and F1-score to assess retrieval performance. For HOVER, we use Macro-F1 scores for claim verification, and F1-score for evidence retrieval.

Baselines As illustrated in Table 1, we compare our proposed method with several advanced baselines to validate effectiveness and superiority. The other baselines treats the two stages independently. BERT Concat (Liu et al. 2020), GAT (Liu et al. 2020), and GEAR (Zhou et al. 2019) modify the ESIM (Chen et al. 2017) model to compute the relevance score between the evidence and the claim for evidence retrieval. KGAT (Liu et al. 2020), DREAM (Zhong et al. 2020), TARSA (Si et al. 2021), Proofver (Krishna, Riedel, and Vlachos 2022), and EvidenceNet (Chen et al. 2022b) utilize BERT-based models trained with a one-hop ranking loss. HESM (Subramanian and Lee 2020) employs a multi-hop evidence retrieval method that is hyperlink-dependent, and GERE is a generative evidence retrieval method. During the reasoning phase, BERT Concat employs a BERT-based model, while Proofver utilizes a seq2seq model to generate natural logic-based inferences as proofs. The remaining methods adopt graph-based approaches.

As depicted in Table 2, we design three types of baselines: BERT-FC (Soleimani, Monz, and Worring 2020) and LisT5 (Jiang, Pradeep, and Lin 2021) are pretrained

Models	Dev		Test	
	LA	FEVER	LA	FEVER
BERT Concat	73.67	68.89	71.01	65.64
GAT	76.13	71.04	72.03	67.56
GEAR	74.84	70.69	71.60	67.10
HESM	75.77	73.44	74.64	71.48
KGAT	78.29	76.11	74.07	70.38
DREAM	79.16	-	76.85	70.60
KGAT+GERE	79.44	77.38	75.24	71.17
TARSA	81.24	77.96	73.97	70.70
Proofver	80.74	79.07	79.47	76.82
EvidenceNet	81.46	78.29	76.95	73.78
TwoWingOS	-	-	75.99	54.33
KGAT+FER	79.02	76.59	73.34	69.61
CD	80.80	78.00	79.30	76.50
MRR-FV	**82.98**	**79.83**	**80.83**	**78.25**

Table 1: Overall performance on FEVER. **Bold** indicates the best result, while underline denotes the second best.

	Models	2-hop	3-hop	4-hop
I	BERT-FC	50.68	49.86	48.57
	LisT5	52.56	51.89	50.46
II	RoBERTa-NLI	63.62	53.99	52.40
	DeBERTaV3-NLI	68.72	60.76	56.00
	MULTIVERS	60.17	52.55	51.86
III	FOLK	66.26	54.80	60.35
	ProgramFC	70.30	63.43	57.74
	MRR-FV	**71.66**	**65.09**	**62.72**

Table 2: Overall performance on HOVER.

models, RoBERTa-NLI (Nie et al. 2020), DeBERTaV3-NLI (He, Gao, and Chen 2023) and MULTIVERS (Wadden et al. 2022) are fact verification fine-tuned models, ProgramFC (Pan et al. 2023) and FOLK (Wang and Shu 2023) are LLM-enhanced models.

Overall Performance

Table 1 presents the performance results of our proposed model MRR-FV on the FEVER dataset, compared to the baselines for fact verification. Overall, MRR-FV demonstrates superior performance across all metrics, underscoring its effectiveness and robustness. It outperforms HESM, indicating that our multi-hop retrieval mechanism can efficiently gather evidence for claim verification without relying on hyperlinks. When compared to the generative retrieval method GERE, MRR-FV shows significant improvement, highlighting the benefits of multi-hop and joint training mechanisms in enhancing the overall performance of fact verification systems. Furthermore, although TwoWingOS, FER, and CD (Fajcik, Motlicek, and Smrz 2023b) utilize the claim verification results to jointly optimize the evidence retriever, they still rely on one-hop methods for evidence retrieval, which fail to capture the interrelationships between evidence pieces during the retrieval process. This limitation is likely the primary reason for their inferior performance compared to MRR-FV.

Models	Dev			Test		
	P	R	F1	P	R	F1
TF-IDF	-	-	17.20	11.28	47.87	18.26
ESIM	24.08	86.72	37.69	23.51	84.66	36.80
BERT	27.29	**94.37**	42.34	25.21	87.47	39.14
XLNet	26.60	87.33	40.79	25.55	85.34	39.33
RoBERTa	26.67	87.64	40.90	25.63	85.57	39.45
GERE	58.43	79.61	67.40	54.30	77.16	63.74
DQN	54.75	79.92	64.98	52.24	77.93	62.55
MRR-FV	**60.37**	87.79	**71.54**	**58.30**	**87.82**	**70.08**

Table 3: Retrieval performance comparison on FEVER.

	Models	2-hop	3-hop	4-hop
One-hop	TF-IDF + BERT	57.2	49.8	45.0
	Oracle + BERT	68.3	71.5	76.4
	CD	81.3	80.1	78.1
Multi-hop	Baleen	81.2	82.5	80.0
	GMR	81.9	82.2	80.2
	MRR-FV	**82.4**	**83.6**	**81.2**

Table 4: Retrieval performance comparison on HOVER.

As depicted in Table 2, following ProgramFC, we divide the HOVER validation set into three subsets based on the number of hops required to verify the claim. On the dev set of HOVER, MRR-FV outperforms the SOTA baseline by 1.9%, 2.6%, and 3.9% on the two-hop, three-hop, and four-hop subsets, respectively. This demonstrates that MRR-FV is more effective for verifying deeper claims.

Performance on Evidence Retrieval

Following previous works (Chen et al. 2022a), we select several representative one-hop evidence retrieval methods as baselines on the FEVER dataset, including TF-IDF (Thorne et al. 2018), ESIM (Hanselowski et al. 2018), BERT (Liu et al. 2020), XLNet (Zhong et al. 2020), and RoBERTa (Zhong et al. 2020). Additionally, GERE (Chen et al. 2022a) is a generative evidence retrieval method, and DQN (Wan et al. 2021) employs a reinforcement learning-based approach to identify precise evidence. In Table 3, we see that MRR-FV's evidence retrieving results mirror its superiority, with large performance gains across the board against the baseline model.

The baselines on the HOVER dataset are divided into two categories: one-hop and multi-hop baselines. As illustrated in Table 4, MRR-FV is capable of retrieving more precise evidence compared to the baselines. Additionally, traditional one-hop methods perform significantly worse than multi-hop methods (Lee et al. 2022), since most claims require synthesizing information from multiple evidence for accurate verification. CD outperforms the other two single-hop baselines as it employs a joint framework.

Models	1-hop	2-hop	3-hop
Traditional multi-hop	0.85	0.76	0.71
Compressor enhanced multi-hop	0.85	0.81	0.79

Table 5: Performance on intention shift mitigation. The evaluation metric is the cosine similarity score between the query at different hop and the original claim

.

Performance on Intention Shift Mitigation

We assess the effectiveness of our proposed method in mitigating the intention shift by evaluating the semantic similarity between query of each hop and the original claim. We benchmark our approach against the "Traditional Multi-hop" baseline, where the query is a concatenation of the original claim and previously retrieved evidence. As depicted in Table 5, as the number of hops increases, the matching degree between the queries of "Traditional Multi-hop" and the initial claim rapidly decreases. This observation aligns with our expectation that concatenating evidence leads to progressively longer queries, which inevitably introduces more noise. In contrast, we leverage a pre-trained query compressor to condense the concatenated sequence while maintaining its semantic integrity. This approach leads to a more gradual decline in the matching degree, indicating that our method can partially mitigate the intention shift.

Ablation Study

As illustrated in Table 6, we design ablation studies to verify the effectiveness of core modules. Specifically, we independently remove or replace a module and observe the impact on the final verification performance.

-w/o Compressor signifies using the evidence retrieved from the previous hops to enhance the claim, without any compression. *-w/o Pre-training* eliminates the pre-training process of the query compressor. *-w/o Multi-hop* retrieves evidence in a single step, without considering the interactions between the evidence during the retrieval process. The evidence retriever of *-w/o Generative Retriever* is replaced by a traditional dense retrieval. *-w/o Intra-sentence Reasoning* eliminates the component that employs the recurrent memory transformer for reasoning within the query, thus neglecting the modeling of long-term dependencies within sentences. *-w/o Inter-sentence Reasoning* processes the claim and evidence directly in a sequential manner, thereby disregarding the modeling of interactions between sentences.

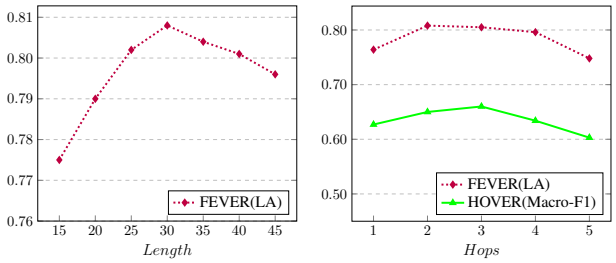
The experimental results show that removing or replacing any module leads to a decline in overall performance, indicating the effectiveness of these modules and their contributes to the enhancement of the overall performance.

Hyperparameter Sensitivity Analysis

As depicted in Figure 3, *Length* and *Hops* dictate the maximum length of the compressed query and the pre-set number of hops for multi-hop retrieval, respectively. The experimental results in Figure 3(a) suggest that both excessively long and overly short compressed text can lead to a decline in performance. Overly long compressed text may not effectively mitigate the dilution of key information in the claim, while excessively short compressed text may fail to preserve the completeness of the evidence. As depicted in Figure 3(b), the evaluation metric on the HOVER dataset is the average Macro-F1 across the three subsets. Similarly, an excessively high or low number of hops can also lead to a decline in performance. Moreover, since the complexity of claims varies across different datasets, each dataset corresponds to a different optimal number of hops. In this regard, our experimental results align with this hypothesis.

	Dev		Test	
Models	LA	FEVER	LA	FEVER
-w/o Compressor	81.94	79.47	80.06	77.58
-w/o Pre-training	82.27	79.62	80.21	77.87
-w/o Multi-hop	81.72	79.32	79.69	77.02
-w/o Generative Retriever	82.13	79.68	80.27	77.82
-w/o Intra-sentence Reasoning	82.49	79.57	80.43	78.15
-w/o Inter-sentence Reasoning	81.86	79.36	79.83	77.37
MRR-FV	**82.98**	**79.83**	**80.83**	**78.25**

Table 6: Performance of ablation study.



(a) Accuracy vs. $Length$ (b) Accuracy vs. $Hops$

Figure 3: Hyperparameter sensitivity analysis.

Conclusion

This paper investigates the novel problem of Multi-hop Retrieval and Reasoning for Fact Verification(MRR-FV), and further propose a generative retrieval enhanced model. We leverage a the Generative Multi-hop Retriever to construct a comprehensive evidence chain by integrating inter-evidence interactions during retrieval, enhanced by a query compressor to mitigate intention shift throughout the multi-hop process. At the reasoning stage, the Hierarchical Interaction Reasoner integrates sequence-based and graph-based approaches, enabling thorough analysis of local inter-sentence correlations alongside global multi-hop features and interrelationships. Finally, experiments on the FEVER and HOVER datasets underscore the superior performance of this method, demonstrating its efficacy in tackling complex fact verification tasks through comprehensive multi-hop evidence retrieval and reasoning.

Acknowledgments

This work was supported by the Natural Science Foundation of China (No.62372057).

References

Bulatov, A.; Kuratov, Y.; and Burtsev, M. 2022. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35: 11079–11091.

Cao, N. D.; Izacard, G.; Riedel, S.; and Petroni, F. 2021. Autoregressive Entity Retrieval. In *The Ninth International Conference on Learning Representations*.

Chen, J.; Kim, G.; Sriram, A.; Durrett, G.; and Choi, E. 2024. Complex Claim Verification with Evidence Retrieved in the Wild. In *The Association for Computational Linguistics: NAACL 2024*, 3569–3587.

Chen, J.; Zhang, R.; Guo, J.; Fan, Y.; and Cheng, X. 2022a. GERE: Generative evidence retrieval for fact verification. In *the 45th SIGIR*, 2184–2189.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced LSTM for Natural Language Inference. In *Proceedings of the Association for Computational Linguistics*, 1657–1668.

Chen, Z.; Hui, S. C.; Zhuang, F.; Liao, L.; Li, F.; Jia, M.; and Li, J. 2022b. EvidenceNet: Evidence Fusion Network for Fact Verification. In *Proceedings of the ACM Web Conference*. ISBN 9781450390965.

Fajcik, M.; Motlicek, P.; and Smrz, P. 2023a. Claim-Dissector: An Interpretable Fact-Checking System with Joint Re-ranking and Veracity Prediction. In *Findings of the Association for Computational Linguistics*, 10184–10205.

Fajcik, M.; Motlicek, P.; and Smrz, P. 2023b. Claim-Dissector: An Interpretable Fact-Checking System with Joint Re-ranking and Veracity Prediction. In *Findings of the Association for Computational Linguistics*, 10184–10205.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

Hanselowski, A.; Zhang, H.; Li, Z.; Sorokin, D.; Schiller, B.; Schulz, C.; and Gurevych, I. 2018. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 103–108.

He, P.; Gao, J.; and Chen, W. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

Hu, X.; Hong, Z.; Guo, Z.; Wen, L.; and Yu, P. 2023. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *the 46th SIGIR*, 2319–2323.

Jiang, K.; Pradeep, R.; and Lin, J. 2021. Exploring Listwise Evidence Reasoning with T5 for Fact Verification. In *The Association for Computational Linguistics: IJCNLP 2021 (Volume 2: Short Papers)*, 402–410.

Jiang, Y.; Bordia, S.; Zhong, Z.; Dognin, C.; Singh, M.; and Bansal, M. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3441–3460.

Khattab, O.; Potts, C.; and Zaharia, M. 2021. Baleen: Robust multi-hop reasoning at scale via condensed retrieval. *Advances in Neural Information Processing Systems*, 34: 27670–27682.

Krishna, A.; Riedel, S.; and Vlachos, A. 2022. ProoFVer: Natural Logic Theorem Proving for Fact Verification. *Transactions of the Association for Computational Linguistics*, 10: 1013–1030.

Lan, Y.; Liu, Z.; Gu, Y.; Yi, X.; Li, X.; Yang, L.; and Yu, G. 2024. Multi-Evidence based Fact Verification via A Confidential Graph Neural Network. *IEEE Transactions on Big Data*.

Lee, H.; Yang, S.; Oh, H.; and Seo, M. 2022. Generative Multi-hop Retrieval. In *EMNLP 2022*, 1417–1436.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Li, C.; Pang, B.; Liu, Y.; Sun, H.; Liu, Z.; Xie, X.; Yang, T.; Cui, Y.; Zhang, L.; and Zhang, Q. 2021. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. In *the 44th SIGIR*, 223–232.

Li, C.; Wang, S.; Yang, D.; Li, Z.; Yang, Y.; Zhang, X.; and Zhou, J. 2017. PPNE: property preserving network embedding. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22*, 163–179. Springer.

Li, C.; Wang, S.; Yu, P. S.; Zheng, L.; Zhang, X.; Li, Z.; and Liang, Y. 2018. Distribution distance minimization for unsupervised user identity linkage. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 447–456.

Liao, H.; Peng, J.; Huang, Z.; Zhang, W.; Li, G.; Shu, K.; and Xie, X. 2023. MUSER: A MUlti-Step Evidence Retrieval Enhancement Framework for Fake News Detection. In *Proceedings of the 29th Conference on Knowledge Discovery and Data Mining*, 4461–4472.

Liu, H.; Wang, W.; Li, H.; and Li, H. 2024. TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection. In *Findings of the Association for Computational Linguistics*, 15556–15583.

Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2020. Fine-grained Fact Verification with Kernel Graph Attention Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7342–7351.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901.

Pan, L.; Wu, X.; Lu, X.; Luu, A. T.; Wang, W. Y.; Kan, M.-Y.; and Nakov, P. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 6981–7004.

Samarinas, C.; Hsu, W.; and Lee, M. L. 2021. Improving Evidence Retrieval for Automated Explainable Fact-Checking. In *The Association for Computational Linguistics: NAACL 2021*, 84–91.

Si, J.; Zhou, D.; Li, T.; Shi, X.; and He, Y. 2021. Topic-Aware Evidence Reasoning and Stance-Aware Aggregation for Fact Verification. In *The Association for Computational Linguistics: IJCNLP 2021*, 1612–1622.

Soleimani, A.; Monz, C.; and Worring, M. 2020. BERT for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval - 42nd European Conference on Research*, volume 12036, 359–366.

Subramanian, S.; and Lee, K. 2020. Hierarchical Evidence Set Modeling for Automated Fact Extraction and Verification. In *EMNLP 2020*, 7798–7809.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *The Association for Computational Linguistics: NAACL 2018*, 809–819.

Wadden, D.; Lo, K.; Wang, L. L.; Cohan, A.; Beltagy, I.; and Hajishirzi, H. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 61–76.

Wan, H.; Chen, H.; Du, J.; Luo, W.; and Ye, R. 2021. A DQN-based approach to finding precise evidences for fact verification. In *The Association for Computational Linguistics: IJCNLP (Volume 1: Long Papers)*, 1030–1039.

Wang, H.; and Shu, K. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6288–6304.

Wu, L.; Wang, L.; and Zhao, Y. 2024. Unified Evidence Enhancement Inference Framework for Fake News Detection. In *IJCAI 2024*, 6541–6549.

Xiao, L.; Zhang, Q.; Shi, C.; Wang, S.; Naseem, U.; and Hu, L. 2024. MSynFD: Multi-hop Syntax Aware Fake News Detection. In *Proceedings of the ACM on Web Conference*, 4128–4137.

Xu, W.; Wu, J.; Liu, Q.; Wu, S.; and Wang, L. 2022. Evidence-aware Fake News Detection with Graph Neural Networks. In *Proceedings of the ACM Web Conference*, 2501–2510.

Yue, Z.; Zeng, H.; Shang, L.; Liu, Y.; Zhang, Y.; and Wang, D. 2024. Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments. In *Proceedings of the 62nd Annual Association for Computational Linguistics (Volume 1: Long Papers)*, 10331–10343.

Zeng, X.; Abumansour, A. S.; and Zubiaga, A. 2021. Automated Fact-Checking: A Survey. *Lang. Linguistics Compass*, 15.

Zhang, C.; Zhang, L.; and Zhou, D. 2024. Causal Walk: Debiasing Multi-Hop Fact Verification with Front-Door Adjustment. In *the 38th AAAI*, volume 38, 19533–19541.

Zhang, L.; Zhang, X.; Li, C.; Zhou, Z.; Liu, J.; Huang, F.; and Zhang, X. 2024a. Mitigating Social Hazards: Early Detection of Fake News via Diffusion-Guided Propagation Path Generation. In *ACM Multimedia*.

Zhang, L.; Zhang, X.; and Pan, J. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *the 36th AAAI*, volume 36, 11676–11684.

Zhang, L.; Zhang, X.; Zhou, Z.; Huang, F.; and Li, C. 2024b. Reinforced adaptive knowledge learning for multimodal fake news detection. In *the 38th AAAI*, volume 38, 16777–16785.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Wang, S.; Philip, S. Y.; and Li, C. 2024c. Early Detection of Multimodal Fake News via Reinforced Propagation Path Generation. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Yu, P. S.; and Li, C. 2025. Knowledge-aware multimodal pre-training for fake news detection. *Information Fusion*, 114: 102715.

Zhang, P.; Guo, J.; Li, C.; Xie, Y.; Kim, J. B.; Zhang, Y.; Xie, X.; Wang, H.; and Kim, S. 2023. Efficiently leveraging multi-level user intent for session-based recommendation via atten-mixer network. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 168–176.

Zhang, X.; and Gao, W. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In *IJCNLP-AACL (Volume 1: Long Papers)*, 996–1011.

Zhao, J.; Li, C.; Wen, Q.; Wang, Y.; Liu, Y.; Sun, H.; Xie, X.; and Ye, Y. 2021. Gophormer: Ego-graph transformer for node classification. *arXiv preprint arXiv:2110.13094*.

Zheng, J.; Zhang, X.; Guo, S.; Wang, Q.; Zang, W.; and Zhang, Y. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. In *IJCAI 2022*, 2413–2419.

Zhong, W.; Xu, J.; Tang, D.; Xu, Z.; Duan, N.; Zhou, M.; Wang, J.; and Yin, J. 2020. Reasoning Over Semantic-Level Graph for Fact Checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6170–6180.

Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2019. GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 892–901.

Zhu, Y.; Si, J.; Zhao, Y.; Zhu, H.; Zhou, D.; and He, Y. 2023. EXPLAIN, EDIT, GENERATE: Rationale-Sensitive Counterfactual Data Augmentation for Multi-hop Fact Verification. In *EMNLP 2023*, 13377–13392.